The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 462–466

ON THE LEAST SQUARES ESTIMATION OF NON-LINEAR RELATIONS

By C. Villegas

Instituto de Matemática y Estadística, Montevideo[1]

1. Introduction. Consider a non-linear relation

$$y = f(x_1, \cdots, x_m; \alpha_1, \cdots, \alpha_p) \tag{1.1}$$

among the real variables $x_1, \cdots, x_m$ and y, where f is a known function and $\alpha_1, \cdots, \alpha_p$ are unknown parameters. The problem of estimating these parameters by least squares methods has been considered recently by Hartley and Booker [1], assuming that the variables x_j are not subject to error and that the variable y is observed with an error which is normally distributed. In this paper, which is only a complement to a previous paper on linear relations [2], these assumptions will be dropped, but it will be assumed instead that replicated observations are available.

2. Notation and model. Suppose that, in order to estimate the non-linear relation (1.1), we have performed an experiment with n replications, which may possibly have an incomplete block design, and suppose that, from the usual statistical analysis of the data, we have obtained k $(m + 1)$ estimators x_{ijn}, y_{in} $(i = 1, \cdots, k; j = 1, \cdots, m)$ converging in probability, when n tends to infinity, to values x_{ij}, y_i which satisfy the non-linear relation (1.1). Using the vector space approach of [2], we shall consider an auxiliary p-dimensional Euclidian space $\mathfrak{U}$ with an orthonormal basis $\mathbf{u}_1, \cdots, \mathbf{u}_p$ and an m-dimensional Euclidian space $\mathfrak{V}$ with an orthonormal basis $\mathbf{v}_1, \cdots, \mathbf{v}_m$ and we shall define on them the linear functionals A, X_{in}, X_i by $A\mathbf{u}_h = \alpha_h$ $(h = 1, \cdots, p)$; $X_{in}\mathbf{v}_j = x_{ijn}$ and $X_i\mathbf{v}_j = x_{ij}$. Then we have

$$y_i = f(X_i, A). \tag{2.1}$$

We shall assume that f is a continuously differentiable function. The value of the differential of f at an arbitrary point $(X, A^*; \Delta X, \Delta A)$ will be denoted by $\Delta X f_X(X, A^*) + \Delta A f_A(X, A^*)$, where $f_A(X, A^*) \varepsilon \mathfrak{U}$ and $f_X(X, A^*) \varepsilon \mathfrak{V}$. We shall write simply

$$\mathbf{f}_i = f_A(X_i, A), \qquad \mathbf{f}_{Xi} = f_X(X_i, A), \tag{2.2}$$

and we shall assume that the vectors $\mathbf{f}_i$ do not lie on any proper subspace of $\mathfrak{U}$. We assume, in addition, that the joint distribution of the $2k$ random variables $n^{\frac{1}{2}}\Delta y_{in}$ (where $\Delta y_{in} = y_{in} - y_i$) and $n^{\frac{1}{2}}\Delta X_{in}$ (where $\Delta X_{in} = X_{in} - X_i$) con-

Received 2 May 1967; revised 29 July 1968.

[1] Now visiting at the University of Rochester. This work has been made with a special support from the Universidad de la República, Montevideo.

verges to a limit distribution (usually a normal distribution, with mean value equal to zero). From this assumption it follows, in particular, by Theorem 3.1 of [2], that the joint distribution of the random variables $n^{\frac{1}{2}}(\Delta y_{in} - \Delta X_{in} f_{Xi})$ converges to the joint distribution of random variables d_i (which, usually, are normally distributed, with mean value equal to zero). We assume also that the expected values $\sigma_{ij} = \mathcal{E}\, d_i d_j$ exist, that the matrix $\{\sigma_{ij}\}$ is positive definite, and that the σ_{ij} are unknown, but consistent estimators s_{ij} are available. Finally, we assume that a preliminary estimator A_n of A is also available. It has been pointed out by a referee that the only condition which must be satisfied by A_n is

$$\|A_n - A\| = O_p(n^{-\frac{1}{2}}). \tag{2.3}$$

In other words, we assume that, given $\epsilon > 0$, there are numbers $K > 0, N > 0$ such that

$$\text{Prob}\,\{n^{\frac{1}{2}}\,\|A_n - A\| \leqq K\} \geqq 1 - \epsilon$$

for all $n \geqq N$. In Section 5 it will be shown how, under mild additional assumptions, a simple preliminary estimator A_n may be obtained.

Sometimes the non-linear relation is given in an implicit form

$$f(x_1, \cdots, x_m; \quad \alpha_1, \cdots, \alpha_p) = 0.$$

This case can be brought back to the model considered in this paper by setting $y_i = y_{in} = 0$.

3. An equivalent linear relation model. Define

$$\mathbf{f}_{in} = f_A(X_{in}, A_n), \tag{3.1}$$

$$g_{in} = y_{in} - f(X_{in}, A_n) + A_n \mathbf{f}_{in}. \tag{3.2}$$

Since f_A is, by assumption, a continuous function, $\mathbf{f}_{in}$ converges in probability to $\mathbf{f}_i$ and therefore g_{in} converges in probability to a limit g_i given by

$$g_i = A\mathbf{f}_i. \tag{3.3}$$

Note that $\mathbf{f}_{in}$, g_{in} are known, but $\mathbf{f}_i$, g_i are unknown. In what follows $\mathbf{f}_{in}$, g_{in} will play the role of the observed data in a linear relation model which we are going to set up. Note also that the "true" values $\mathbf{f}_i$, g_i satisfy the linear relation (3.3) whereas the "observed" values $\mathbf{f}_{in}$, g_{in} typically will not satisfy it. Define the "error" e_{in} by

$$g_{in} = A\mathbf{f}_{in} + e_{in}. \tag{3.4}$$

Obviously e_{in} converges in probability to zero. From (3.2) and (3.4) we have

$$e_{in} = y_{in} - f(X_{in}, A_n) + \Delta A_n \mathbf{f}_{in}, \tag{3.5}$$

where $\Delta A_n = A_n - A$. By the differentiability of f at the point (X_i, A), and by our hypothesis about ΔX_{in} and ΔA_n we have

$$e_{in} = \Delta y_{in} - \Delta X_{in} \mathbf{f}_{Xi} + o_p(n^{-\frac{1}{2}}),$$

and consequently the joint distribution of the random variables $n^{\frac{1}{2}}e_{in}$ converges to the joint distribution of the random variables d_i.

4. Results. If, in addition, we assume that, with probability 1, the matrix $\{s_{ij}\}$ is positive definite and the vectors $\mathbf{f}_{in}$ do not lie on any proper subspace, then the hypotheses made in [2] are satisfied, and we may apply the estimation theory developed there. In order to write the results in the notation of [2], consider an auxiliary k-dimensional Euclidian space $\mathfrak{L}$, with an orthonormal basis $\mathbf{w}_1, \cdots, \mathbf{w}_k$ and define the linear transformations F_n, $F\colon \mathfrak{L} \to \mathfrak{U}$ and the linear functionals G_n, G by

$$F_n\mathbf{w}_i = \mathbf{f}_{in}, \qquad F\mathbf{w}_i = \mathbf{f}_i,$$

$$G_n\mathbf{w}_i = g_{in}, \qquad G\mathbf{w}_i = g_i.$$

From the theory developed in [2] it follows that, if the linear transformation $S\colon \mathfrak{L} \to \mathfrak{L}$ is defined by $(S\mathbf{w}_i, \mathbf{w}_j) = s_{ij}$, then the S-least squares estimator, defined by

$$\tilde{A}_n = G_nS^{-1}F_n'(F_nS^{-1}F_n')^{-1}, \tag{4.1}$$

where $F_n'\colon \mathfrak{U} \to \mathfrak{L}$ is the adjoint of F_n, is asymptotically efficient within the class of ordinary estimators associated with the model (3.4). In the usual case in which the d_i are normally distributed, with mean values equal to zero, the error of prediction $(\tilde{A}_n - A)\mathbf{u}$, for any fixed $\mathbf{u} \in \mathfrak{U}$, is asymptotically normally distributed, with asymptotic mean value equal to zero and asymptotic variance equal to the inner product $(\mathbf{u}, (F\Sigma^{-1}F')^{-1}\mathbf{u})$, where Σ is the limit of S. Note that, although the data of the linear model (3.4) depend on the preliminary estimator which is used, the minimum asymptotic variance of an efficient estimator is independent from it. It will be convenient to say that an *ordinary estimator for the non-linear relation* is an ordinary estimator for an associated linear model like (3.4), corresponding to any admissible preliminary estimator. Clearly, then, the least squares estimator $\tilde{A}_n$ defined by (4.1) is asymptotically efficient within the whole class of ordinary estimators for the non-linear relation, in the sense that it minimizes the asymptotic mean square error of prediction within that class.

In the case in which the deviations e_{in} are asymptotically independent, with asymptotic mean square errors $\sigma_{ii} = \sigma_i^2$, we can take as S the linear transformation defined by $S\mathbf{w}_i = s_i^2\mathbf{w}_i$, where s_i^2 is a consistent estimator of σ_i^2. In this case the S-least squares estimator is the linear transformation which minimizes the weighted sum of squares of deviations

$$\sum s_i^{-2}(g_{in} - A\mathbf{f}_{in})^2. \tag{4.2}$$

As was pointed out by a referee, the assumption that, with probability 1, the matrix $\{s_{ij}\}$ is positive definite and the vectors $\mathbf{f}_{in}$ do not lie on any proper subspace may be dropped without any serious consequences. In effect, in such a case, *in probability*, that is, with a probability which tends to 1 when n tends to infinity, the matrix $\{s_{ij}\}$ is positive definite and the vectors $\mathbf{f}_{in}$ do not lie on any

proper subspace. It can be easily seen that this is all that is required in order that all the above mentioned results remain valid, provided that the ordinary estimators of a linear relation be redefined in the following way: a random linear functional A_n^* is an ordinary estimator of A, if $A_n^* = G_nL_n$, where $L_n: \mathfrak{U} \to \mathfrak{L}$ is a random linear transformation such that, in probability, F_nL_n is equal to the identity transformation.

It can be proved also that the following definition is equivalent: a random linear functional A_n^* is an ordinary estimator of A, if, in probability,

$$A_n^* = G_nC_nF_n'(F_nC_nF_n')^{-1}, \tag{4.3}$$

where $C_n: \mathfrak{L} \to \mathfrak{L}$ is a random positive definite transformation. The equation (4.3), which was conjectured by a referee, implies, as was pointed out by a second referee, that the notion of ordinary estimator is closely associated with the notion of projection. In effect, $C_n^{\frac{1}{2}}F_n'A'$ is the projection of $C_n^{\frac{1}{2}}G_n'$ over the range of the linear transformation $C_n^{\frac{1}{2}}F_n'$ (see p. 1681 of [2]).

5. A preliminary estimator. Assume that the equations

$$y_i^* = f(X_i^*, A^*) \qquad (i = 1, \cdots, p), \tag{5.1}$$

have a unique solution

$$A^* = \varphi(X_1^*, \cdots, X_p^*; \quad y_1^*, \cdots, y_p^*)$$

in a neighborhood of the point $(X_1, \cdots, X_p; y_1, \cdots, y_p)$, and that φ is continuous at this point. Then

$$A_n = \varphi(X_{in}, \cdots, X_{pn}; y_{1n}, \cdots, y_{pn}) \tag{5.2}$$

is defined in probability and is a consistent estimator of A. We shall prove now that A_n is an admissible estimator, i.e., that it satisfies the condition (2.3). Obviously A_n minimizes

$$Q_n(A^*) = \sum_{i=1}^{p} [y_{in} - f(X_{in}, A^*)]^2. \tag{5.3}$$

Assuming that the set of possible values of the parameter is open, by differentiation of $Q_n(A^*)$ we have, whenever A_n is defined,

$$\sum [y_{in} - f(X_{in}, A_n)]\mathbf{f}_{in} = 0. \tag{5.4}$$

By substitution of (3.5) in (5.4) we have

$$\sum_{i=1}^{p} \mathbf{f}_{in}\, \Delta A_n \mathbf{f}_{in} = O_p(n^{-\frac{1}{2}})$$

and, if F_0 is the restriction of F to the subspace spanned by $\mathbf{w}_1, \cdots, \mathbf{w}_p$

$$\|\Delta A_n[F_0F_0' + o_p(1)]\| = O_p(n^{-\frac{1}{2}}).$$

If we assume that $\mathbf{f}_1, \cdots, \mathbf{f}_p$ do not lie on any proper subspace, then by Theorem 3.5 of [2] F_0F_0' is invertible and

$$\|\Delta A_n\| \leqq \|(F_0F_0')^{-1} + o_p(1)\|\, O_p(n^{-\frac{1}{2}}),$$

from which (2.3) follows immediately.

6. Comparison with other least squares estimators. The sum of squares (4.2) may be written also as

$$\sum s_i^{-2}[y_{in} - L_{in}(A^*)]^2, \tag{6.1}$$

where

$$L_{in}(A^*) = f(X_{in}, A_n) + (A^* - A_n)f_A(X_{in}, A_n) \tag{6.2}$$

is the first order Taylor development of $f(X_{in}, A^*)$. Hence, if A_n is a good preliminary estimator, then $\tilde{A}_n$ will be near the estimator $\hat{A}$ which minimizes

$$Q_n(A^*) = \sum_{i=1}^{k} s_i^{-2}[y_{in} - f(X_{in}, A^*)]^2,$$

and which will be called the *non-linear least squares estimator*. In the hope of getting estimators which are closer to $\hat{A}_n$, it has been recommended to follow the Gauss-Newton iteration procedure, in the jth step of which we minimize (6.1) taking as preliminary estimator the estimator obtained in the previous step. However, we clearly see now that the *iterated least squares estimators* obtained in this way are also ordinary estimators, and therefore all of them have the same asymptotic efficiency as the estimator obtained in the first step of the iteration procedure, and consequently, as far as the asymptotic efficiency is concerned, there is no need to proceed further with the iteration method. This does not mean, however, that the small- and moderate-sample size properties cannot be improved by proceeding further with the iteration, and additional research, using perhaps Monte Carlo methods, would be desirable to clarify this point.

We shall compare now our estimator $\tilde{A}_n$ with the non-linear least squares estimator $\hat{A}_n$. We shall assume that $\hat{A}_n$ is a consistent estimator of A, since otherwise any further comparison is unnecessary. By an argument similar to that made in the previous section, it may be proved that, if $\hat{A}_n$ is consistent, then it satisfies the condition (2.3). Hence we can use $\hat{A}_n$ as a preliminary estimator in order to compute an asymptotically efficient $\tilde{A}_n$ minimizing (6.1). By differentiation of (6.1) we obtain

$$\sum [y_{in} - f(X_{in}, \hat{A}_n)]f_A(X_{in}, \hat{A}_n) = 0.$$

Therefore, (6.1) is equal to

$$\sum [y_{in} - f(X_{in}, \hat{A}_n)]^2 + \sum [(A^* - \hat{A}_n)f_A(X_{in}, \hat{A}_n)]^2.$$

Since this is obviously minimized by $\tilde{A}_n$, it follows that $\hat{A}_n$ is also an ordinary least squares estimator, and therefore, as far as the asymptotic efficiency is concerned, there is no need to choose $\hat{A}_n$ instead of the far simpler to compute estimator $\tilde{A}_n$.

REFERENCES

[1] HARTLEY, H. O. and BOOKER, AARON (1965). Nonlinear least squares estimation. *Ann. Math. Statist.* **36** 638–650.

[2] VILLEGAS, C. (1966). On the asymptotic efficiency of least squares estimators. *Ann. Math. Statist.* **37** 1676–1683.